View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by E-LIS

Relazione tecnica su CDSware

Antonella De Robbio
[documento 21-10-2002]

Premessa

La presente relazione nasce da un incontro avuto con i tecnici del CERN che hanno dato la loro piena disponibilità ad illustrarci lo strumento CDSware.
L'incontro si è tenuto presso il CERN, la mattina precedente l'inizio dei lavori del Workshop Europeo "*The Open Archives Initiative (OAI): Gaining Independence with e-prints archives and OAI*", svoltosi al CERN di Ginevra dal 17 al 19 ottobre 2002.

Presenti all'incontro:

- **Informatici CERN**: Jean-Yves Le Meur, Thomas Byron, Martin Vesely, Tibor Simko

- **Gruppo italiano** composto da bibliotecari, informatici e un docente di biblioteconomia: Antonella De Robbio (Università di Padova), Valentina Comba e Simone Sacchi (Università di Bologna), Silvana Mangiaracina e Alessandro Tugnoli(CNR di Bologna), Alberto Salarelli (Università di Parma)

Il gruppo italiano ha posto una serie di domande, preventivamente predisposte, che hanno riguardato i seguenti punti:

1. Requisiti per l'installazione
2. Capacità di trattare grosse quantità di dati e di operare in un contesto complesso come un ateneo
3. Quali sono le differenze tra la versione del 1° agosto e la prossima del 1° novembre
4. L'organizzazione degli archivi, sistema dei filtri o collezioni pre-organizzate
5. Integrazione tra database catalografico (gestionale) e CDSware: sistema di gestione dei dati, formato dei record …
6. Motore di ricerca: caratteristiche, configurazioni, modalità di ricerca, indicizzazione dei dati
7. L'architettura: parti componenti e moduli
8. OpenURL e SFX in CDSware per il reference linking
9. Configurazione multilingue
10. Compatibilità con lo Z39.50
11. Export dei dati, formati: XML; UNIMARC o MARC
12. Gestione, manutenzione, assistenza anche in relazione alle risorse umane

Dalla discussione durata circa due ore, sono emerse una serie di considerazioni utili al lavoro che si sta portando avanti negli atenei italiani (e CNR). In particolare, per quanto riguarda Padova la riflessione sullo strumento CDSware riguarda sia il gruppo GLACS (Analisi e Comparazione Sistemi di Automazione) sia il gruppo GLIRE (Integrazione Risorse Elettroniche).
In fondo alla presente relazione annoto alcune considerazioni.

Introduzione a CDSware

CDSware (CERN Document Server Software) http://cdsware.cern.ch è un insieme di moduli software, in soluzione OAI compatibile, per la gestione di collezioni di dati, sviluppato dai tecnici del CERN che consente la ricerca incrociata su set di archivi differenti. CDSware è un software libero e viene rilasciato gratuitamente sotto licenza di tipo GPL GNU.
E' interamente costruito con software libero: Apache, MySQL, PHP, Python, WML in 80.000 linee di codice. Tutte le personalizzazioni sono basate su web, come pure l'amministrazione del server.
CDSware offre ampie possibilità di personalizzazione: interfaccia configurabile come portale atta ad ospitare varie tipologie di collezioni, potente motore di ricerca con sintassi simile a Google-like, personalizzazioni utente, incluse document baskets e email notification alerts, sottomissione elettronica dei documenti e caricamento di vari tipi e formati di documenti.
Implementa funzioni sia di data provider sia di service provider ed è in grado di scambiare metadati tra eterogenei repositories.
Consente la raccolta e l'indicizzazione di metadati dai formati più eterogenei, sia in modalità batch sia tramite il Protocollo OAI per l'Harvesting dei Metadati, offredno un'interfaccia unica di consultazione per più repositories.
Usa MARC21 come standard bibliografico interno.
La prima versione è stata rilasciata il 1° agosto 2002 e si compone di un singolo package contenente tutti i moduli (tra cui "submit" e "search"), una prossima versione verrà rilasciata il prossimo 1° novembre e conterrà le applicazioni per i moduli "convert" e "agendas").
Vi sono due mailing list disponibili, la prima relativa alle novità, la seconda per le discussioni tra gli implementatori.

I servizi offerti da CDSware

I servizi a disposizione degli utenti sono i seguenti, resi disponibili da una serie di moduli o parti componenti che verranno descritti in seguito.

SEARCH - http://weblib.cern.ch/
: Permette la ricerca entro le informazioni bibliografiche del catalogo e dentro al testo pieno dei documenti. Offre numerose personalizzazioni all'utenza.

SUBMIT - http://doc.cern.ch/Submit
: Permette la sottomissione elettronica dei documenti

CONVERT - http://doc.cern.ch/Convert
: Offre possibilità di convertire I documenti caricati dalgi utenti da un foramto ad un altro(es. da MS Word a PDF).

SCAN - http://doc.cern.ch/Scanning
: Offre possibilità di scansione dei documenti cartacei (questo servizio è solo interno alla rete intranet del CERN e per ora non è incluso in CDSware).

AGENDA - http://agenda.cern.ch/
: Offre possibilità di pianificare meetings o workshops con i dettagli sui singoli interventi.

Il contesto CERN in cui è nato CDSware

Due parole sull'attuale contesto in cui CDSware si è sviluppato e in cui è applicato, per comprendere meglio la portata dello strumento.
Le aree dei documenti del CERN si riferiscono alla fisica e discipline correlate, matematica, astrofisica, …, ma è ovviamente indifferente la copertura disciplinare sulla quale CDSware opera.
Numerose sono le collezioni su cui CDSware è applicato. Il software consente infatti visualizzazioni dedicate alle attività dei singoli gruppi di ricerca o di attività, consente il recupero e la visualizzazione dei documenti da collezioni speciali come quelle degli esperimenti, delle divisioni e così via.
CDSware è in grado di operare in tempi rapidissimi, su una vastissima gamma di collezioni e documenti. Nella sua applicazione al CERN infatti serve:

- 156.000 host differenti solo nel 2001
- 17.000 differenti hosts/clients per mese
- 1.000 consultazioni (visite) al giorno
- 3.500 ricerche vere e proprie dalle sue interfacce al giorno
- 50.000 "hits" e 1.5 GB di traffico di rete giornaliero

CDSware opera in sostanza sui seguenti dati:

- 350 diverse collezioni
- 565.000 record
- 220.000 full-text

I documenti trattati sono in prevalenza preprint, nell'ordine di circa 1000 ingressi a settimana.
Questo dimostra la scalabilità del software in rapporto al trattamento di grosse quantità di dati e la sua flessibilità in rapporto al recupero (harvesting) di documenti da fonti diverse.
Al CERN va detto si cataloga solo il 5% del materiale, il resto viene tutto recuperato per il 70% dal noto ArXiv, e per il restante 25% da altri archivi e banche dati.

L'architettura di CDSware

Le parti componenti di CDSware riguardano le seguenti configurazioni:

1. **WebSubmit: Submitting data**
2. **BibHarvest: harvesting OAI repository**
3. **BibConvert: harvesting non-OAI collections**
4. **BibFormat: Formatting and linking records**
5. **WebSearch: Searching metadata/citations/full text**
6. **BibWord: Indexing metadata and full text**
7. **WebAccess: Managing complex collection hierarchy**
8. **WebPerso: Personalizing web access**
9. **BibData: Modifying records (librarians only)**

1. WebSubmit, l'OpenArchive integrato: Sottomissione dei dati, conversione automatica dei formati, generazione del numero automatico di report, funzioni multiple per la post-sottomissione.

Questa componente riguarda la **sottomissione (deposito) dei dati**, da parte degli autori.Per dati si intendono sia il full-text dei documenti, sia i relativi metadati che descrivono il documento proposto per

la sottomissione. La sottomissione esclusivamente via Web e pu essere effettuata dagli stessi autori (auto-archiviazione), o da personale di segreteria o dallo staff dei bibliotecari. La procedura per la sottomissione avviene per step successivi: apertura, monitoraggio, approvazione (peer reviewing) con controlli ad ogni step.
In questa componente esiste il modulo per la **conversione automatica di documenti**, da un formato all altro. Si tratta di un software componente molto potente (del tipo quello offerto da Tom server della Carnegie Mellon University) in quanto numerosi sono i formati di partenza accettabili e numerose le possibilit di una loro conversione in altro formato.
Altra caratteristica della componente WebSubmit è la **generazione automatica del numero di report** e timbratura digitale del numero sul documento.
Una serie di funzionalità relative ai processi di post-sottomissione caratterizzano questo modulo come OpenArchive o server di e-print multifunzionale e integrato nei processi di ricerca da una parte e nei servizi bibliotecari dall'altra.
In particolare vengono offerte le seguenti funzionalità:

- -alerting (per il singolo utente) e possibilità di una ridistribuzione dei dati a liste di discussione (utilissimo per i gruppi di ricerca)
- -capacità di una gestione dei commenti da parte dei pari (peers)
- -possibilità di modificare i metadati sottomessi
- -invio delle versioni full-text revisionate (non solo della lettera all'autore)
- -estrazione delle citazioni (per l'inclusione in apposito database citazioni dove si può agire con un sistema di reference linking)
- -estrazione della lista autori, se lunga (questo aspetto riguarda principalmente i lavori dei fisici degli esperimenti)
- -estrazione di parole chiave (funzione in via di definizione)

Sostanzialmente questa parte componente consente di avere un e-print server o OpenArchive dentro CDSware, integrato con tutte le altre funzioni (OPAC, raccolta dati …)

2. **La strategia di harvesting di CDSware**

I due moduli **BibHarvest** e **BibConvert** permettono di eseguire importazioni massicce di record da fonti diverse, attraverso template per la descrizione delle fonti da cui eseguire il caricamento o per descrivere la trasformazione (modifica) delle fonti stesse.

- ❑**BibHarvest:** Harvesting dagli archivi OAI
  E' il modulo che si occupa della raccolta (harvesting) di dati (in questo caso metadata) dai repository OA compatibili, dai data provider
- ❑**BibConvert**: Harvesting da collezioni non OAI compatibili
  Con questo modulo è possibile raccogliere record da collezioni non OAI compatibili

I dati raccolti sono sempre convertiti in un formato MARC XML OAI compatibile, utilizzato al CERN come formato di rappresentazione interna dei dati.
Va detto che questi moduli sono in grado di recuperare anche i documenti (full-text) e non solo i metadati.

3. La strategia di linking di CDSware

Il modulo **BibFormat** è adibito alla formattazione flessibile e connessione dei record (linking)
Tutte le informazioni di linking sono separate dale informaizoni bibliografiche. In altri termini il metadato non contiene il link diretto alla risorsa a testo pieno, ma le informazioni utili alla connessione risiedono fuori dal metadato, nell'ottica del protocollo OpenURL.

BibFormat supporta due differenti tipi di risoluzione di link:

-link esterni: genera il collegamento “al volo” in base alle regole immagazzinate per quella risorsa

-link interni il link è sempre un file, il sistema controlla se esite, l’accesso, i formati …

4. La ricerca

La ricerca avviene attraverso un potente e veloce motore simile a Google.
E’ inoltre implementato il protocollo OAI (versione 2.0).
Il formato di rappresentazione dei dati è il MARC21 il quale consente ad ogni campo di essere ricercabile e browsable singolarmente.
E’ inoltre supportata la ricerca full-text anche tra i seguenti formati: PostScript, PDF, Msword, MSExcel, MSPowerPoint.
I metadati, le citazioni e il full-text possono essere ricercabili assieme attraverso operatori booleani.
Tutte le opzioni di ricerca possono essere personalizzate in ordine a:

-campi su cui eseguire la ricerca
-ordinamento dei record
-formati dei record: HTML breve o dettagliato, XML OAI DC+MARC21, …
-risultati suddivisi per collezione, anche con gerarchie complesse

Esiste la possibilità di personalizzare da parte degli utenti alcune funzioni: il layout, il sistema di alerting e il basket con il proprio scaffale dei “preferiti”, …
I moduli coinvolti sono i seguenti:

- **WebSearch:** Funzionalità di ricerca e meta-ricerca
  Questo modulo può essere considerato come un OPAC evoluto o meglio un SuperOPAC che consente di ricercare attraverso funzionalità molto spinte sia nei metadati, sia nelle citazioni, sia nel full-text dei documenti. Le interfacce di ricerca sono numerose e personalizzabili e quindi WebSearch quale SearchEngine è in grado di ricercare in tutte le collezioni presenti, e non solo nel gestionale (catalogo).
- **BibWord:** Indicizzazione dei dati
  Questo modulo si occupa dell’indicizzazione dei dati in connessione con il modulo WebSearch. Crea gli indici a partire dai metadati e dallo stesso full-text al fine di un recupero dell’informazione più appropriato alle query poste.
- **WebAccess:** E’ il modulo che consente di trattare gerarchie complesse, come le viste sui soggetti per un browsing espanso a più livelli
- **WebPerso**: Personalizzzazione dell’accesso web
- **BibData**: Modifica records (solo per lo staff dei bibliotecari)
  Questa funzionalità consente ai bibliotecari dello staff di operare modifiche sui metadati.
  Si tratta comunque di uno strumento non raffinato che non consente la catalogazione completa, infatti CDSware deve comunque appoggiarsi ad un software gestionale per la catalogazione bibliografica completa.

Riassumendo, l’architettura di CDSware risulta la seguente:

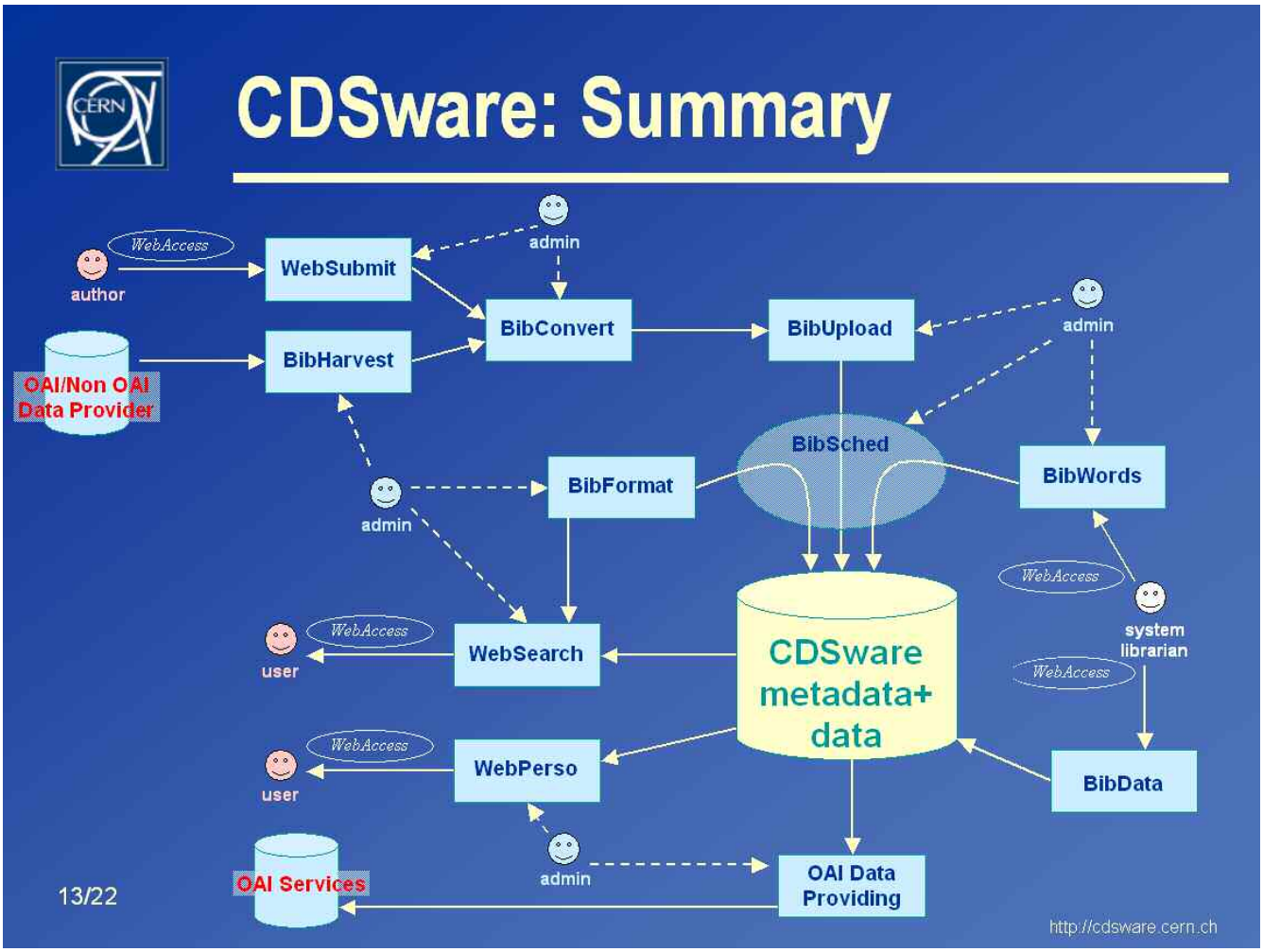

immagine tratta dalle slide di presentazione di Jean-Yves Le Meur [1]

Considerazioni finali

CDSWare fa parte della categoria di prodotti come MetaLib o iPort, che potremmo definire "portali o gateway per l'integrazione di risorse" o strumenti OpenACces.
La sua estensibilità lo rende uno strumento in grado di interagire con qualsiasi altra applicazione o installazione specifica. Altri servizi su piattaforma indipendente possono essere integrati in CDSware, giacché il server di conversione è come un servizio aggiuntivo offerto all'interno della installazione stessa.
Non si tratta di un gestionale, ma di uno strumento che consente una visibilità e ricercabilità del catalogo gestionale in integrazione con le altre collezioni o archivi del sistema informativo.
Può essere applicato a qualsiasi sistema gestionale o base di dati catalografica, tenendo conto che lavora in MARC21 e quindi opportune modifiche andrebbero operate per la conversione in UNIMARC.
Non è z39.50 compatibile.
Esiste solo in versione inglese e quindi dovrebbe essere tradotto in lingua italiana, considerato che è uno strumento rivolto soprattutto all'utenza finale. Attualmente non è dotato di opzione multilingua.

I requisiti tecnici sono una semplice macchina Linux o Unix che supporti Apache, MySQL, PHP, Python, WML.

Contatti

Il CERN offre pieno supporto nell'installazione e configurazione del prodotto, sia a distanza sia inviando personalmente un tecnico sul luogo (servizio a pagamento).
Il prodotto come detto sopra è gratuito.
La presenza di un tecnico CERN risulterebbe inoltre assai utile per la formazione di personale locale.

◆CERN Document Server
•http://cds.cern.ch/
◆CDSware sources, mailing lists, demo
•http://cdsware.cern.ch/
◆Contact
·cds.support@cern.ch

Riferimenti bibliografici

1. Jean-Yves Le Meur, Building OAI repository with the CERN Document Server Software
http://documents.cern.ch/cgi-bin/setlink?base=agenda&categ=a02333&id=a02333s11t1/transparencies

2. Martin Vesely , CERN Document Server Software
http://www.oaforum.org/workshops/pisa_abstracts.php

3. Hector Sanchez, Flexelink: the new CDS Link Manager
http://documents.cern.ch/cgi-bin/setlink?base=agenda&categ=a02335&id=a02335s1t0/transparencies